An Interlingua-Based Chinese-English MT System

QI Xuan (齐 璇), ZHOU Huiping (周会平) and CHEN Huwang (陈火旺)

School of Computer Science, National University of Defense Technology, Changsha 410073, P.R. China
E-mail: qixuanhao@hotmail.com
Received January 15, 2001; revised September 24, 2001.

Abstract Chinese-English machine translation is a significant and challenging problem in information processing. The paper presents an interlingua-based Chinese-English natural language translation system (ICENT). It introduces the realization mechanism of Chinese language analysis, which contains syntactic parsing and semantic analyzing and gives the design of interlingua in details. Experimental results and system evaluation are given. The result is satisfying.

Keywords machine translation, syntactic parsing, semantic analysis, interlingua

1 Introduction

Machine translation is a significant and challenging problem in information processing. The Chinese language is used by the largest number of people in the world. English is the language that is used most popularly. So, Chinese-English and English-Chinese machine translation tools are very important. Now there are many English-Chinese MT systems whereas few Chinese-English MT systems could be seen on the market. One of the reasons is that analyzing Chinese is very difficult and few Chinese-English MT systems are mature.

Several groups of different universities did some research in MT technology. Harbin Institute of Technology developed a Chinese-English MT system in which the pattern matching method was used[1]. Northeastern University did much research in Chinese semantic analysis[2,3]. Tsinghua University used dependency grammar in Chinese parsing[4]. There are HuanYuTong and SinoTrans Chinese-English MT systems on the market. HuanYuTong was the first one tested in 1998. Its translation speed is 7 characters per second and its readability rate is up to 80 percent. But it faces many problems in practice. Its translation quality is not satisfying yet and its translation speed is slow.

The CLIP laboratory of Maryland has conducted a feasibility study for fully automatic, high quality and interlingua-based Chinese-English translation for the newswire text in the economic domain[5]. The interlingua is based on language-independent lexical conceptual structure (LCS) representations. The study used the idea of the Unitran[6-8] system which emphasized semantic representations. KANTOO[9] is an object-oriented C++ implementation of CMU’s KANT technology for machine translation. KANT[10-12] uses controlled source languages, and explicit yet focused semantic models for each technical domain to achieve very high accuracy in translation.

Our group has designed and developed an interlingua-based Chinese-English natural language translation system — ICENT, which includes four main modules that are Chinese segmentation and tagging, Chinese analysis, converting Chinese parsing results to inner interlingua forms and English generation. Because Chinese lacks modality, it cannot be analyzed perfectly only by syntactic parsing. Semantic analyzing is stressed in our system. The paper presents the structure of ICENT and elaborates the realizing mechanism of ICENT. Section 2 presents the structure and goal of ICENT. Section 3 presents the design of ICENT. Section 4 gives the experimental result and system evaluation. And Section 5 gives the conclusion.

The project is supported by the National ‘863’ High-Tech Programme of China (Grant No.2001AA114110).
2 The Architecture and Goal of ICENT

ICENT is an interlingua-based system. First, Chinese sentences are segmented and tagged to get information about words. Second, the results of the above steps are parsed to produce syntactic and semantic structures. Third, syntactic and semantic structures are converted to interlingua representations. Last, target language sentences are generated from the interlingua. The flow chart is shown in Fig.1.

The primary goal of our project is to implement an interlingua-based Chinese-English natural-language translation system, which processes uncontrolled Chinese sentences. We selected four Chinese textbooks of elementary school for grade one and grade two. These four textbooks cover almost all sentence types although they contain not so many words. The detailed goals of ICENT are shown below:

- Uncontrolled source language and syntax;
- Without manual pre-editing;
- Fully automatic parsing and disambiguating;
- High-speed on-line translation;
- Fault tolerance;
- Extensibility.

3 The Design of ICENT

ICENT has four modules shown in Fig.1. Chinese text parsing and the interlingua design are the keys to our project. The parsing combines syntactic parsing and semantic analyzing so that more accurate representations can be got. The interlingua combines syntactic knowledge and semantic knowledge, which can represent almost all sentence types. The details are described below.

3.1 The Segmentation and Tagging of ICENT

Chinese text is a character sequence without any space so the first step of Chinese-English MT is segmentation and tagging. The step is important and the result will influence the quality of Chinese parsing directly. Peking University developed an algorithm\cite{13} for segmentation and tagging, which could reach an accuracy of 90 percent. We have adopted the algorithm and improved it for subsequent semantic analysis.

There are two segmentation and tagging dictionaries, one is for the system and the other is for users. Users can add unknown words to the dictionary. The segmentation is done with the rule of maximum matching and at the same time each word is tagged with possible parts of speech. There
may be segmentation ambiguity and tagging ambiguity so syntactic rules and statistic knowledge are used for disambiguation.

We conduct not only syntactic tagging but also semantic tagging. Semantic class tagging is done on syntactic tagging results and each word may have several semantic classes. Disambiguation will be done with semantic analysis. Thus what we get is a POS and semantic class sequence.

3.2 The Parsing of ICENT

3.2.1 Characteristics of Chinese

Chinese grammar is very flexible. Some word classes can act as different syntactic components without any modal change. A syntactic component can act as different parts of speech. A Chinese sentence often has several verbs and each verb has not any special modal character. How to determine which verb is the main one is a very important and difficult problem in Chinese parsing. The ellipsis phenomenon is very common in Chinese. The structure of Chinese phrases is almost the same as that of Chinese sentences. Among phrases semantic restrictions are much stronger than syntactic restrictions. So, ICENT combines semantic analysis with syntactic parsing. With syntactic parsing semantic analysis is done also and at the same time syntactic structures and semantic representations are derived.

3.2.2 Syntactic Parsing

The phrase is a rather steady syntactic unit in Chinese. We can analyze the structure of a sentence by analyzing the grammar relations among the components during the reduce process. There are some approaches of parsing. Harbin Institute of Technology used the pattern matching method[1] and Tsinghua University used the dependency grammar in Chinese parsing[4]. What we used is the extended context-free grammar (ECFG)[1][4], which is extended by grammar functions and precondition judge functions. ECFG is a high efficiency grammar and the grammar is defined below.

\[ G = (V_T, V_N, V_D, V_F, S, \psi) \]

where

- \( V_T \) is a finite set of terminal symbols and it is not empty;
- \( V_N \) is a finite set of nonterminal symbols, it is not empty, and \( V_T \cap V_N = \emptyset \), and
- \( V_D \) is a finite set of grammar functions and it is not empty. Grammar functions describe the grammar relations among components during the reduce process. Each production has a corresponding grammar function.

\( V_F \) is a finite set of precondition judgment functions, it is not empty, and \( \text{accept} \in V_F \) and \( \text{NULL} \in V_F \). Precondition judgment functions guide the actions of the analysis program. In precondition judgment functions, semantic analyzing will be done, which will study the content of the current stack and judge whether the current operation of the analyzing process is reasonable. According to the judgment, some wrong analysis trees can be aborted in time and representations can be produced for correct analysis trees. NULL is the empty function, which means accepting unconditionally. Accept is the function that judges whether the current analysis tree is reasonable when analyzing is accomplished. Each production has a precondition judgment function.

\( S \) is the start symbol and \( S \in V_N \).

\( \psi \) is a production set. The form of a production is \( P \rightarrow \langle \alpha, \beta, \gamma \rangle, P \in V_N, \alpha \in (V_T \cup V_N)^*, \beta \in V_D, \gamma \in V_F \). \( \beta \) describes grammar relations among the components of \( \alpha \). \( \gamma \) is the precondition judgment function. A production is effective only if the return value of the precondition judgment function is true.

In the context-free grammar, two productions are the same only if the left of one production is the same as the left of the other and the right of one production is the same as the right of the other. In our extended context free grammar, these two productions are not the same if their grammar functions are not the same. This means that two phrases may have the same hierarchy, but different grammar relations. So when we get a sentence’s structure, we can get the grammar relations among
the components of the sentence at the same time. These grammar relations can eliminate its structural ambiguity in analysis trees, which can make the sentence structure much clearer.

The conventional LR algorithm can only manage simple context-free grammar that has no ambiguity. But ambiguity is one of the main characteristics of natural languages and the production set of the extended context-free grammar has ambiguity too. This means that the ACTION table and the GOTO table generated by the production set have ambiguities too. The ambiguity of the ACTION table means that a term of the ACTION table may include shift action and reduce action at the same time or includes several reduce actions. The ambiguity of GOTO table means that a certain state of a stack can go to several states with a certain symbol. We have improved the LR algorithm, making it handle shift-reduce ambiguity. When meeting shift-reduce ambiguity, the parsing stack will be copied as many times as the number of actions in the current parsing table. Thus each parsing stack can do a unique action. Using this method, the analyzer can get all possible states and keep all possible results unless disambiguation is done.

3.2.3 Semantic Analysis

Through syntactic parsing only, what we get is the surface structure of a sentence. Because Chinese is quite different from English and lacks modality, it is difficult to generate a correct natural target language only by the surface structure of the source language. Deep structures are helpful for generation and this needs semantic analysis. In syntactic parsing, all possible results are kept. The number of the results may be very large although the correct result may be only one. We have to do disambiguation for the parsing results to improve accuracy and efficiency. Not only syntactic information but also semantic information are needed to do so.

Case grammar[12] describes semantic relations among noun phrases and verb phrases by semantic roles but it could not describe the difference of the main verb and other verbs in Chinese and could not describe semantic relations among the components of a noun phrase or a verb phrase. Frame description[16,17] is a good method describing the hierarchy of a sentence. We innovated the case grammar to describe Chinese semantic relations more expediently and used frame description to describe Chinese.

We designed a Chinese semantic hierarchical system[18–20]. Just like syntactic units including word, phrase and sentence, semantic units include semantic primitive, sense and semantic chunk. A semantic primitive is the basic unit of the semantic hierarchy, which describes the basic concept. A sense represents a concept, which is described by primitives. A word may have several senses. A semantic chunk describes semantic relations of a compound concept. A semantic chunk may correspond to either a phrase or a sentence. The formal definition of semantic units is shown below.

SemUnit ::= semantic primitive | sense | semantic chunk
Sense ::= name[note]
Semantic chunk ::= name (HEAD SemUnit)
(SUBJECT subjectframe)
(OBJECT objectframe)*
(EVENT eventframe)*
(COMPLEMENT compframe)*

subjectframe ::= (SEMROLE role) objectframe ::= (SEMROLE role)
(SEMCORE SemUnit)) (SEMCORE SemUnit))

eventframe ::= (SEMROLE role) compframe ::= (SEMROLE role)
(SEMCORE SemUnit)) (SEMCORE SemUnit))

name ∈ ΨMSP, ΨMSP ⊆ ΨSP, ΨMSP is the set of main semantic primitives, ΨSP is the set of semantic primitives.
note ∈ ΨSP, note describes the secondary semantic feature of a sense.
role ∈ ΨSR, ΨSR is the set of semantic roles.
A semantic chunk represents the semantic relations among components that constitute the compound concept. It is expressed with the case frame, which has a frame name and a frame body. The frame name describes the main semantic feature of the semantic chunk. The frame body is composed of some slot-values. The value of the slot HEAD is the semantic unit, which is the main component of a semantic chunk. When the main component is an event, the value of the slot SUBJECT is a subject frame (subjectframe), which describes the logical subject of the semantic chunk. The value of the slot OBJECT is an object frame (objectframe), which describes the logical object of the semantic chunk. The value of the slot EVENT is an event frame (eventframe), which describes other events besides the main event. The value of the slot COMPLEMENT is a complement frame (complementframe), which describes the complement of the main component. The subjectframe contains two slot-values, the value of the slot SEMROLE is the semantic role that the logical subject acts, the value of the slot SEMCORE is the semantic unit that acts as the logical subject. The definitions of the objectframe, the eventframe and the complementframe are similar to the definition of the subjectframe.

Semantic roles express semantic relations among components of a semantic chunk. A semantic chunk is defined as a nested frame, which is useful for semantic reduction together with syntactic reduction. Semantic structures of different levels can be generated with corresponding syntactic structures.

For example, the Chinese sentence ‘小/ (xiao3/“little”) 山羊 (shan1yang2/“goat”) 吃 (/chi1/“eat”) 青草 (/qing1ciao3/“green grass”)’ has the semantic structure as follows:

```
(HEAD eat | 吃 (HEAD Sense(吃))))
(SUBJECT (livestock| 牲畜
  ((SEMROLE AGENT)
    (SEMCORE livestock| 牲畜
      ((HEAD Sense(山羊))
        (MODIFIER size| 尺寸 (HEAD Sense(小)))))))))
(OBJECT (FlowerGrass| 花草
  ((SEMROLE PATIENT)
    (SEMROLE FlowerGrass| 花草 (HEAD Sense(花草)))))))
```

‘小山羊’ is the agent of the event ‘吃’, ‘青草’ is the patient of the event and ‘小’ is the modifier of the thing ‘山羊’.

Senses can constitute a semantic chunk and chunks can constitute a larger semantic chunk. How senses constitute a semantic chunk and how chunks constitute a larger semantic chunk are defined by semantic rules.

Semantic rules are the abstract of semantic knowledge, which describe the semantic relations among senses, between senses and chunks, among chunks and so on. A semantic rule is composed of two parts, a well-formed semantic link and a generated chunk template. A well-formed semantic link is a concept sequence that satisfies logical restraints. Each element of the sequence contains the main semantic feature of the concept (name_i) and a restrictive condition (c_i). A generated chunk template is the abstract of the semantic chunk generated by the well-formed semantic link. The semantic role that each element of the well-formed semantic link acts is given in the template and the role value is not a special semantic unit, but the element position in the link. The definition of a semantic rule is shown below.

Semantic Rule ::= Well-formed Semantic Link → Semantic Chunk Template
Well-formed Semantic Link ::= name_0/c_0 . . . name_n/c_n
where name_i ∈ N_MSP and c_i is a restrictive condition.

With semantic rules semantic chunks can be constituted. For example, the main semantic feature ‘size| 尺寸 /di3ch3m4/’ is an appearance characteristic, livestock has the size characteristic, so size| 尺寸 + livestock| 牲畜 is a well-formed semantic link. Thus ‘小’ and ‘山羊’ can constitute a semantic chunk ‘小/山羊’ because ‘小’ has the main semantic feature of ‘size| 尺寸’ and ‘山羊’ has the main
semantic feature of ‘livestock’ 牲畜’. The new semantic chunk has a main semantic feature too, which is assigned by rules.

With semantic rules semantic analysis can be done. The semantic analysis is done with the syntactic parsing at the same time. In syntactic parsing, a production can be carried out only if the components accord with a semantic rule. This means that if the current syntactic components cannot satisfy any semantic rule, the current syntactic tree is wrong and should be discarded. Thus at the time of syntactic parsing, semantic analysis is done too and when getting a syntactic structure, a semantic structure is also obtained. In the same way, we can use well-formed semantic links to select the correct sense of a word.

We developed a tool to get semantic rules. Because the amount of semantic primitives is more than one thousand, the number of semantic rules that we have got is very large. We should abstract and refine these semantic rules.

3.3 The Interlingua of ICENT

ICENT is an interlingua-based translation system, all the results of the source language analysis are stored in the interlingua and all the information that the target language generation needs is fetched from it. How to design the interlingua is the core of the system.

UNITRAN and KANT are two well-known interlingua-based MT systems. The interlingua of UNITRAN[^8] is based on a theory of lexical semantic description. The semantic representation is useful for analysis but the conceptual grain size seems too small. The interlingua of KANT[^11,12,21,22] is based on syntax. Because KANT is used for controlled language, the conceptual grain size is large and it cannot describe uncontrolled language freely.

We have compared the differences between Chinese and English, designed and implemented a kind of interlingua, which combines syntactic and semantic descriptions. The design of the interlingua follows the pragmatic principle. The content of the interlingua is complemented and improved in the process of the design of the system. The method of iterative refinement can make it satisfy the needs of Chinese analysis and English generation gradually.

The structure of the interlingua is a recursive framework[^23]. A framework is composed of one pair or many pairs of slot-values. These slot-value pairs describe the attributes of the framework. Slots represent the types of attributes and values represent the values of attributes. For example, the pair (CAT, ADJ) means the word class of the object is an adjective. CAT is the name of a slot, it means what the slot represents is a word class. ADJ is the value of CAT. All the information of the interlingua is expressed by using these slot-value pairs.

Each frame has a name that represents the object expressed by the frame. Now the biggest frame unit is the sentence frame, which can be a simple sentence frame or a compound sentence frame. The compound sentence frame is composed of subsentence frames. Relations among subsentences are expressed in each subsentence. The simple sentence frame contains attributes, which are TENSE, ASPECT, MOOD and so on, and predicate frames. The definition of the sentence frame is shown below.

\[
\text{sentence-frame} := \langle \text{compoundsent-frame} \rangle \mid \langle \text{simplesent-frame} \rangle
\]

\[
\text{compoundsent-frame} := \langle \text{pattern compound} \rangle
\text{type} \langle \text{value} \rangle
\text{adverbials} \langle \text{adverbials-frame} \rangle
\text{subsent} \langle \text{subtype} \langle \text{value} \rangle \rangle
\text{conjunction} \langle \text{c-structure} \rangle
\text{sent} \langle \text{sentence-frame} \rangle
\ast
\]

\[
\text{simplesent-frame} := \langle \text{pattern simple} \rangle
\text{mood} \langle \text{value} \rangle
\text{tense} \langle \text{value} \rangle
\text{aspect} \langle \text{value} \rangle
\]
(CONJUNCTION (c-structure))
(PREDICATE (predicate-frame)))

The predicate frame is composed of one head frame, one subject frame, one or two object frames, one or several subpredicate frames and one or several adverbial frames. The head frame describes the structure of the main verb. The subject frame describes the logical subject, which contains the semantic role of the logical subject and the structure of it. The object frame describes the logical object, which contains the semantic role and the structure too. The subpredicate frames describe predicates except the main predicate. The type of subpredicate frames includes RESULTEVENT, ACCOMPANIMENT, SUCCEEDING and so on. The adverbial frames describe the components that modify a sentence. The type of adverbial frames includes TIME, LOCATION, MANNER and so on.

We give the definition of the subject frame as an example.

(subject-frame) := ((SUBTYPE (subjecttype-value))
(SUBCORE (component-frame)))

(subjecttype-value) := AGENT | RELEVANT | EXPERIENCER

The smallest frame unit is the c-structure, which describes a word that has a certain sense. The c-structure includes the syntactic and semantic attributes of a word. Its structure is shown below.

(c-structure) := NIL | ((CAT cat-value)
(ROOT word)
(SEM semantic-value)
(NUMBER number-value)
(MEASURE word)
(SELF (bool-value))
(MEAREPEAT (bool-value)))

CAT denotes the word class, ROOT denotes the morphology, NUMBER denotes the quantity, MEASURE denotes the measure word, MEAREPEAT denotes the repetition of the measure word, SELF denotes reflexive and SEM denotes the semantic definition. The c-structure is the most basic unit of the interlingua and all the frames can be got by adding slot-value pairs to it.

The interlingua of ICENT is nested, which can express Chinese sentences simply and naturally. The syntactic and semantic structures produced by Chinese analysis are similar to the interlingua, so converting from syntactic and semantic structures to the interlingua representation is easy and convenient. The interlingua includes the syntactic and semantic information, by which English sentences can be generated easily. Now the interlingua is used in practice and it can describe a mass of Chinese phenomena.

3.4 English Generation of ICENT

The interlingua structure of a sentence is the deep structure, and the generation module can get syntactic and semantic information from it to generate an English sentence. The generation module finds the corresponding English syntactic structure for each sub-frame of the interlingua of the sentence, converts the interlingua frame to the English syntactic structure, generates each syntactic structure, and generates each English sentence with the information of the sentence frame. The syntactic information in the interlingua provides the correct grammar description for English sentence generation, while the semantic information in the interlingua provides the correct logical order for English sentence generation.

Because of the difference between Chinese and English, some information needed by English generation cannot be got easily from Chinese analysis. For example, it is difficult to generate the article and number for English nouns. Context information can make it easier, but the translation unit of ICENT is sentence and the context information is not enough to do so. We will study the application of pragmatics in MT in the future.
4 Experimental Result

We did an experiment using the Chinese textbooks of elementary school for grade one and grade two. Although the vocabulary of these textbooks is not so large, syntactic and semantic phenomena are very complex. The experiment is run on a PC with Pentium III, 550Hz and 128M memory. We translated 1,461 sentences using ICENT. There are 12,919 characters that are equal to 8,900 words. The total time of translation is 59 seconds. That means the translation speed is up to 219 characters per second.

We invited three doctoral students of computer science to read the translation results. If what a student understands when he reads a target English sentence is equal to the meaning of the source Chinese sentence, we think the sentence can be understood and is acceptable. On the contrary, we think the sentence cannot be understood and is unacceptable. The result is shown in Table 1 below.

<table>
<thead>
<tr>
<th></th>
<th>No. of unacceptable sentences</th>
<th>No. of acceptable sentences</th>
<th>Acceptable rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first person</td>
<td>415</td>
<td>1046</td>
<td>72%</td>
</tr>
<tr>
<td>The second person</td>
<td>218</td>
<td>1213</td>
<td>83%</td>
</tr>
<tr>
<td>The third person</td>
<td>172</td>
<td>1289</td>
<td>88%</td>
</tr>
<tr>
<td>Average acceptable rate</td>
<td></td>
<td></td>
<td>81%</td>
</tr>
</tbody>
</table>

Suppose the sentences that are translated compose set $I$, the sentences that the first student cannot understand compose set $A$, the sentences that the second student cannot understand compose set $B$ and the sentences that the third student cannot understand compose set $C$. NUM($X$) is the number of the members of set $X$. Then NUM($I$) = 1,461, NUM($A$) = 415, NUM($B$) = 248, NUM($C$) = 172, NUM($A \cap B$) = 195, NUM($B \cap C$) = 142, NUM($A \cap C$) = 146, NUM($A \cap B \cap C$) = 123, NUM($A \cup B \cup C$) = 475. Let $D = A \cap B \cap C$, then no one of the three persons can understand the sentences of set $D$. This means that at least one person can understand the sentences of set $I - D$. Thus 92% of the sentences of set $I$ can be understood by at least one person. Let $E = A \cup B \cup C$, then at least one person can understand the sentences of set $E$. This means that every one can understand the sentences of set $I - E$. Thus 67% of the sentences of set $I$ can be understood by every person. The acceptable rate range is from 67% to 92%.

We also used Huanyutong in translating these sentences. The translation speed is 11 characters per second. The acceptable rate is 82%.

From the experimental results, we can conclude that the speed of our system is high and the translation quality is almost the same as other systems.

5 Conclusion and Future Work

This paper presents the design and implementation of an interlingua-based Chinese-English natural language translation system (ICENT). Especially, the paper presents the method of parsing, which includes syntactic parsing and semantic analyzing, and gives the design of the interlingua.

We have developed a PC-based prototype system. The translation results from the testing corpus are satisfying. We have generated a syntactic-semantic electronic dictionary that has nearly 50,000 words. Next we should expand the dictionary to cover more words. We have developed a tool to get semantic rules from the testing corpus. Now we are implementing an algorithm to abstract semantic rules automatically. To get a more natural target language, we have much work to do with the English generation.

In conclusion, there are many things to do in the future. With more effort, we believe that ICENT will get out of the laboratory and become a practical system.
References


Qi Xuan is a Ph.D. candidate in the School of Computer Science, National University of Defense Technology (NUDT). Her research interests include machine translation and computational linguistics.

Zhou Huiping received his Ph.D. degree from the School of Computer Science, NUDT in 2000. His research interests include machine translation and artificial intelligence.

Chen Huowang is a professor in the School of Computer Science, NUDT. His research interests include computer software and artificial intelligence. He is a member of the Chinese Academy of Engineering.